Speech quality in oral oropharyngeal cancer patients

de Bruijn, M.J.

2013

document version
Publisher's PDF, also known as Version of record

Chapter 5

Speech quality in patients treated for oral or oropharyngeal cancer: validation of objective speech analyses

5

Marieke de Bruijn
Louis ten Bosch
Birgit I. Witte
Johannes A. Langendijk
C. René Leemans
Irma Verdonck- de Leeuw

Submitted

Abstract

Speech quality in patients treated for oral or oropharyngeal cancer is most often evaluated subjectively but can also be assessed objectively by analyses of speech recordings. In earlier studies, we developed and tested various acoustic-phonetic (AP) and artificial neural network (ANN) analyses separately. The present study aims to validate these objective speech quality analyses in the same patient cohort and in a new patient cohort (external validation). Speech quality was evaluated subjectively regarding hypernasality, articulation, intelligibility and by patient reported speech outcome. AP analyses were performed on vowels /a, i, u/, stop consonants /k, p, b, d, t/ and fricative /x/. ANN analysis of the feature nasalance was performed on /a, i, u/ and on the entire stretch of speech; ANN analysis of the feature voicing was performed on /p, b, d, t/. In patient cohort 1 Intelligibility was predicted by AP analysis of /p/ and vowel space and by ANN analysis of /d/. Articulation was predicted by AP analysis of vowel space and by measurements of the feature 'voicing' for /b/. Nasal resonance was predicted by AP analysis of /a/, /x/ and /b/. Patient reported speech outcome was predicted by AP analysis of /i/ and /k/ and by ANN analysis of /p/. The amount of variance explained varied from moderate to. In cohort 2 Intelligibility was predicted by AP analysis of /a/, /i/ and /x/. Articulation was predicted by AP analysis of vowel space and by measurements of the feature 'voicing' for /p/. Nasal resonance was predicted by AP analysis of /p/ and /t/. Patient reported speech outcome was predicted by AP analysis of /u/ and /t/ and by ANN analysis of /d/. The amount of variance explained varied from moderate to poor. Objective analyses of speech quality in HNC patients are valid and contribute moderately to a multidimensional speech evaluation protocol.

Introduction

Speech quality is one of the tumour specific quality of life domains that is often compromised in patients treated for oral or oropharyngeal cancer.[1-3] Speech problems occur in 40-70% of these patients, caused by the relatively large amounts of tissue that are damaged or removed in the oral cavity or oropharynx. In addition, radiotherapy may cause fibrosis and stiffening of the tissues of the vocal tract. Surgical tumour resection is associated with a lower intelligibility and worse articulation also after reconstructive surgery. Patients are less able to quickly and correctly produce speech sounds which negatively affect intelligibility and communicational suitability. Speech problems often lead to impaired social functioning and lower quality of life.[4-6]

Perceptual evaluations by professionals such as speech therapists or patient reported outcome through questionnaires are most often used to assess speech quality in clinical practice.[4, 7-9] Perceptual speech quality assessment protocols do not allow to wide sharing between different groups of physicians or speech therapists in various hospitals. Moreover, despite elaborate attempts to design perceptual rating instruments for speech quality, one has to conclude that it is only feasible to obtain high consensus on broad aspects of speech such as tempo and intelligibility but that listeners do not succeed at attaching reliable judgments on more specific aspects of speech, such as nasality or articulation.[10, 11] Therefore, there is an urgent need for an objective speech quality assessment tool to be used in clinical practice and for research purposes.

Three previously performed pilot studies in the first stage of the development of objective speech analyses in patients treated for oral or oropharyngeal cancer, revealed several speech sounds distinguishing patients from controls.[5, 12, 13] Speech sounds were analysed through acoustic phonetic (AP) analyses and through the use of an artificial neural network (ANN).[1] The vowel space regarding the cardinal vowels /a, i, u/, and amount of hypernasality, the stop consonants /p, t, b, d/, and the velar speech

[1] This is an analysis based on a representation as obtained by automatic estimation of articulatory features, performed by means of an artificial neural network.

sounds /k, x/ were identified as potential speech quality markers that categorizes deviant speech as produced by patients from normal speech as produced by healthy controls. Also, within patients these speech sounds differentiated with regard to tumour stage: patients treated for larger tumours having worse speech quality. However, the results of the three pilot studies revealed that the predictive value is moderate. This result may be explained by the fact that the three studies were performed separately, and it is hypothized that better results could be obtained when all variables are combined into one model.
The goal of the present study is to combine the results from the previous studies and to externally validate the findings onto another cohort of patients treated for oral or oropharyngeal cancer.

Patients and methods

Patients
All patients were treated by surgery and postoperative radiotherapy at the Department of Otolaryngology–Head & Neck Surgery and the department of Radiation Oncology of the VU University Medical Center in Amsterdam, the Netherlands, between 1994 and 2003. Patients underwent composite resections for advanced oral or oropharyngeal squamous cell carcinoma with microvascular soft tissue transfer for the reconstruction of surgical defects. Surgery consisted of composite resections including excision of the primary tumour with en bloc ipsilateral or bilateral neck dissection. In case of oropharyngeal carcinomas, a paramedian mandibular swing approach was used. All free flaps were successful. Patients received radiotherapy in case of advanced (T3–4) tumours, positive or close surgical margins, multiple lymph node metastases and/or extra nodal spread. The primary site received a dose of 56 to 66 Gy in total (2 Gy per fraction, 5 times per week), depending on surgical margins. The nodal areas received a total of 46, 56 or 66 Gy (2 Gy per fraction, 5 times a week) in case of N0, N+ without extranodal spread and N+ with extranodal spread, respectively. Exclusion criteria were incapability to participate in functional tests and difficulty communicating in Dutch. Written informed consent was obtained from all patients.

Patient cohort 1 included 51 patients at 6 months after treatment and aged 23 to 73 years (mean: 53.8 years, sd: 8.7 years). Patient cohort 2 included 64

patients at six months to nine years after treatment (median: 1.5 years, range: 8.5 years) and aged 26 to 87 years (mean: 60.4 years, sd: 10.8 years). See table 17 for additional information concerning gender and clinical parameters.

Table 17. Characteristics of 51 patients in cohort 1 and 64 patients in cohort 2.

	Gender		Tumour site		T-classification		(chemo-) Radiation therapy	
	Male	Female	Oral cavity	Oro-Pharynx	2	3-4	Yes	No
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Cohort 1	28 (55%)	23 (45%)	21 (41%)	30 (59%)	26 (51%)	25 (49%)	47 (92%)	4 (8%)
Cohort 2	35 (55%)	29 (45%)	31 (48%)	33 (52%)	41 (64%)	23 (36%)	43 (67%)	20 (31%)

Speech recordings

Patients read-aloud a standardized Dutch text. The distance between lips and microphone was 30 centimeters. Speech recordings were conducted in a sound attenuated booth. For each speaker the recording level was adjusted to optimize signal-to-noise ratio. The recorded speech was digitized with Cool Edit PRO 1.2 (Adobe Systems Incorporated, San Jose, CA, USA) with 22-kHz sample frequency and 16-bit resolution.

Subjective speech evaluation

Perceptual evaluation of speech quality comprised ratings of intelligibility, articulation and hypernasality by two speech pathologists on the entire stretch of running speech. To enable subjective speech evaluation, a computer program was developed to perform blinded randomized listening experiments and to store intelligibility, articulation and hypernasality scores in a database. The panel of trained listeners rated articulation and hypernasality on a 4-point scale, ranging from normal to increasingly deviant speech quality. Intelligibility was scored using a 10-point scale, following the Dutch educational grading system where 1 represents the worst score and 10 represents the best score and 6 is just sufficient. Interrater agreement for subjective assessment of intelligibility ranged from 40% to 90%. Intrarater agreement for two repeated speech fragments of articulation and hypernasality was high with 100% equal scores between the ratings.

Speech problems in daily life as reported by patients were assessed by the Speech Subscale (including 3 items) of the EORTC Quality of Life Questionnaire H&N35 module.[14, 15] The scores are linearly transformed to a scale of 0 to 100, with a higher score indicating a higher level of speech problems. A detailed description of these subjective speech ratings and results can be found in Borggreven et al.[4]

Objective speech evaluation

In the present study, vowels /a, i, u/, velar consonants /k, x/ and stop consonants /b, d, p, t/ were objectively analysed by acoustic-phonetic analyses and by artificial neural network analysis. See table 18 for an overview of speech analyses specifications.

Table 18. Overview of speech material and Acoustic-Phonetic (AP) and Artificial Neural Network (ANN) analyses.

	Acoustic-Phonetic Analyses	Artificial Neural Network
/a, i, u/	Formant 1 Formant 2 Vowel space	Feature 'nasal'
/x/	Spectral slope	-
/k/	Burst percentage	-
/b, d, p, t/	Duration VOT + burst	Feature 'voicing'
entire text	Speaking rate (words/minute)	Feature 'nasal'

- Acoustic-phonetic analyses

Speaking rate of the entire stretch of speech was calculated in words per minute.

Of the vowels /a, i, u/ the first formant frequency (F1) and the second formant frequency (F2) were studied in the present study. Vowels are – compared to consonants – relatively easy to identify in the speech signal and to analyze acoustically. Vowel formant analyses proved to be valid measures of speech quality in patients with deviant speech originating from oral cancer or other origins in earlier studies.[8, 16] Vowel identity (or its spectral color) is characterized by acoustic correlates and is primarily determined by its formants. Broadly speaking, F1 is associated with 'height': the degree of opening of the vocal tract. F2 is associated with the anterior-posterior tongue position.[17] Plotting the vowels /a, i, u/ onto a graphical F1-F2 representation shows the vowel space (more specifically the vowel space) (see figure 9). The vertices of the vowel space represent the most extended positions. The area of the vowel space is a measure for the amount of reduction in the vowel system and can (formally) be measured in terms of Hz^2.[18]

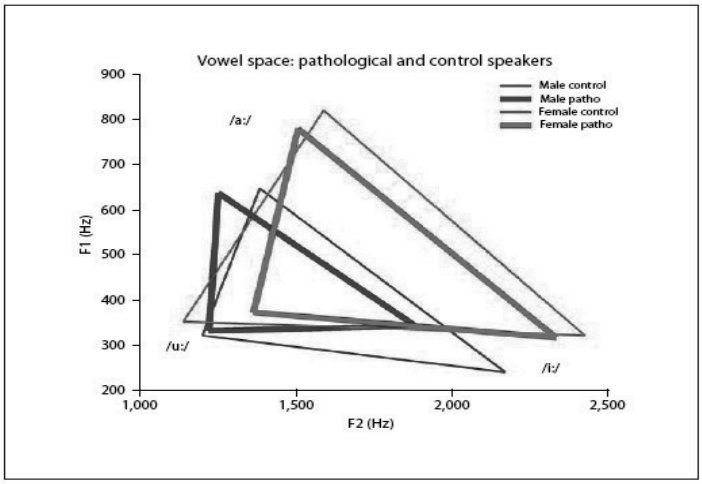

Figure 9. Vowel space of male (blue) and female (pink) patients (fat lines) and of controls (thin lines).

The velar consonants /k/ and /x/ were analyzed because earlier research revealed that patients with an oral or oropharyngeal tumour often have difficulties with the production of velar speech sounds: speech raters often mistook /k/ for /x/.[4, 19] For /k/ the duration of air pressure release (the so-called plosive) as a percentage of the total duration was measured (see figure 10: calculations on /k/ are comparable to those on /t/). For /x/ the spectral slope was used as outcome measure. It describes how quickly the amplitude decreases concurrently with increase of frequency. The spectral slope is of influence on sound quality and timbre and is used for speech discrimination and voice recognition.[20]

Of stop consonants /b, d, p, t/ the duration of voice-onset-time (VOT) and burst were measured. In normal speech, VOT is one of the functionally most relevant parameters that distinguishes voiced and voiceless stops and is a result of the temporal coordination of voicing and oral articulation gestures. VOT is defined as the length of time that passes between when a stop consonant is released and when voicing, the vibration of the vocal folds, begins. For voiced consonant stops /b, d, g/, the voicing starts before the burst of airflow. This voice sound preceding the burst in voiced stop consonants is typical for the Dutch language. For voiceless consonant stops /p, t, k/, a short period of silence precedes the burst. Voiced plosives /b, d, g/ usually have a shorter VOT than voiceless plosives /p, t, k/ (see figure 2). VOT is significantly related to speaking rate.[17, 21-23] The voiced stop consonants /b/ and /d/ and their voiceless counterparts /p/ en /t/ were used as speech material (in the Dutch phonological system the voiced counterpart /g/ of the voiceless stop consonant /k/ is not present and therefore these velar stop consonants were not investigated). Duration of VOT and the duration of the following release of air pressure were measured. Where this paper refers to the VOT, actually the pre-burst silence portion is meant.

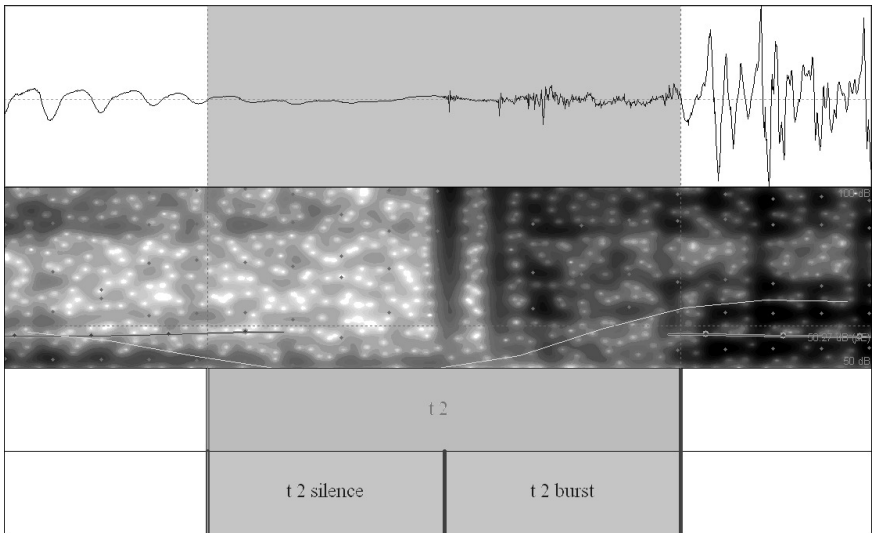

Figure 10. Example of segmentation of /t/. The upper panel tier shows the waveform of the entire duration of /t/ ("t 2"); the second panel displays the spectrogram while the third panel contains the segmentation and labelling partition in silence ("t 2 silence") and burst ("t 2 burst"). The spectrogram shows the relative absence of voicing during the pressure-building silence, followed by the outburst of voicing during the burst. This figure is made in Praat.[24]

Since the acoustic realization of certain speech sounds may depend on its context, we took different phonological contexts around the target speech sounds into account, in order to improve generalization. For each selected speech sound two acoustic realizations were segmented from running speech and were acoustic- phonetically analyzed using the speech processing software Praat.[24] A spectrogram functioned as a visual representation of the speech signal, which facilitated recognition of phonemes in the speech signal and facilitated precise extraction of phonemes from running speech. Spectral and acoustic speech analyses were automatically performed using scripts.[24]

- Artificial neural network and feature analysis

In the context of research on Automatic Speech Recognition (ASR), several speech decoding techniques have been developed that are able to automatically segment and label an unknown utterance in terms of phone-like segments.[25, 26] In the last decade however, new approaches were designed that circumvent the use of phone symbols in the segmentation of speech. Instead of using phonetic symbols (mostly predefined by the ASR development of the ASR system), these methods focus on more basic properties of the speech signal that characterize the speech signal in a way more similar to acoustic features or phonological features such as 'manner' and 'place' of articulation.[27, 28] In the last decade, several classification techniques have become available, such as Artificial Neural Nets (ANNs) and Support Vector Machines.[29-31]

In the present study we used ANNs to obtain an articulatory feature representation of an input speech signal. ANNs contain a number of model parameters (weights of the connections between network nodes) that determine the relation between input and output of these ANNs. The parameters can be trained on a training set in which input and desired output of the ANN are specified for each training data point. The ANNs used in this paper had an input context consisting of 7 consecutive mel-frequency cepstral coefficients (MFCC) frames, which are mathematical coefficients for sound modelling. During training these were used in combination with a canonical value (0 or 1) of the phonological feature that was being modelled by the ANN. The ANNs had been trained on speech from speakers without reported articulatory problems and in the test provided estimations of phonological features from the acoustic input signal. The ANNs used in the present study were trained on the basis of speech originating from the corpus of the Institute of Phonetic Sciences, University of Amsterdam, the Netherlands (IFA Spoken Language Corpus). This corpus contains speech in a variety of styles of four normal speaking males and four normal speaking females in combination with accompanying phoneme labels.[32] For the training of ANN, speech spectra of one healthy male and one healthy female in combination with accompanying segmentations were derived from the IFA-corpus. Several normalisation steps applied during the training of the ANN classifier (including mean and variance normalisation of the MFCCs on utterance level, energy normalisation) led to the best possible generalization

performance of the eventual trained ANNs. ANNs were specifically trained on labelled speech spectra for all phonological features.[33]

Training took place via error-back propagation, one of the well-known and commonly used training procedures for ANNs. After training, ANN-voicing was tested on speech of two other speakers from this IFA corpus. The output of ANN-voicing varies between 0 (absent) and 1 (present). A value of 0.8 for instance means that voicing is rather strongly present in the speech frame. Depending on the amount of input and of consistency in labels during training, ANN-voicing achieves high levels of correct classification. In the quality assessment of ANNs, the performance is given in terms of frame accuracy. Performance during testing was 80% correct at frame level. An accuracy of 80% means that the classifier correctly classifies the feature value assigned to this frame for 80% of all frames in the evaluation test set. The degree of the phonological feature voicing as identified by the artificial neural network was determined for VOT as well as for the following burst of the selected stop consonants.

In the experiments described here, ANNs were used to estimate the degree of various phonological features such as manner, place, voicing, front-back and rounding. Each of these properties was modelled by an ANN. Each ANN was modelled by a three layer feed-forward network: one input layer, one hidden layer, and one output layer. The input layer is fed with the MFCCs obtained from the MFCC extraction step. The units in the output layer represent the estimated values of the various options for that particular feature. For example, the manner feature is modelled by the manner-ANN which has 6 units on its output layer (see also chapter 1, table 1):

Manner: 0/NULL-approx-fric-nasal-stop-vowel

These six units in the manner-ANN estimate the degree of NULL, approximant, fricative, nasality, stop, and vowel, for each 10ms frame in the input speech signal, respectively. The NULL value is a unit that takes positive values if the network is not able to positively assign values to any of the other five units or when that makes no sense such as during silence. In total, the manner-ANN provides 6 values for each frame of 10 ms.

By taking into account the output of the other five ANNs (place, voicing, front-back, rounding, and static) and stacking all results, we obtain a 28-dimensional feature vector for each frame of 10 ms. In the present study, we particularly focus on the phonological features voicing (ANN-voicing) and nasality (ANN-nasal), that is, on specifically two out of this collection of 28 values per 10ms. One of the most important model parameters in an ANN is the dimension (number of hidden units) of the hidden layers. In this study, we adopted a setting that has been suggested in the literature[29] and that showed good results in training and test sessions with speech from healthy controls.

The ANNs that were used in this study are available as public domain software.[34] In the present study we used phonological features voicing and nasal (in the sequel, the resulting ANN is referred to as ANN-voicing and ANN-nasal).The motivation to use ANNs instead of Support Vector Machines (SVM) is determined by the facts that compared to SVM, ANNs deliver a relatively small model and ANNs use continue mapping. The model does not encounter discrete selections during the classification task, as is the case with SVM.

Statistical analysis

Pearson correlation coefficients were used to investigate univariate associations between the objective analyses on the one hand and subjective speech evaluations of intelligibility, articulation, nasal resonance and patient-reported outcome at the other. In case of significant differences of values of the two realisations of each speech sound, both realisations were used separately. In case of no differences of the values of the two realisation of each speech sound, the average of the two realisations was used. Univariate correlations with a p-value $<.20$ were also inserted into the multivariate regression analyses, meaning that only correlations indicating a (significant) coherence were used for further statistical analysis into predictive value of subjective evaluations.

Multivariate regression analyses were performed to obtain insight into the role of objective parameters in predicting subjective speech evaluation. For intelligibility and patient reported speech outcome stepwise linear regression was used, while for articulation and nasal resonance conditional forward

logistic regressions were performed on a binary scale [normal (score 0) vs. deviant (scores 1-3)].
This procedure was performed in cohort 1 and then repeated -as external validation- in cohort 2. As a result hereof, the variables may vary in the final models belonging to subjective (self-)evaluations in cohort 1 and cohort 2.

Results

Univariate Pearson correlations
For cohort 1, both positive and negative Pearson correlations (table 19a) revealed that intelligibility was significantly related to AP analyses of vowel space, /i, u/, /k/ and /b, p/. Subjective evaluation of articulation was significantly related to AP analyses of vowel space, /i/and to ANN-voicing in/d/. Subjective evaluation of nasality was significantly related to AP analyses of /a, i/ and /x/ and to ANN-nasal in /i/. Patient reported speech outcome was not significantly related to AP analyses and not to ANN-voicing nor ANN-nasal.

Table 19 a. Cohort 1: Pearson correlation coefficient r between objective speech parameters and subjective parameters: intelligibility, articulation, nasality and patient-reported speech outcome. "/a/1" means the first realisation of the two /a/'s that are included in the study, "F1" is the first formant, "F2" is the second formant, "VOT" is the voice-onset-time of stop consonants, "ANN" is analysis by an artificial neural network. Phonemes are "averaged" when there is no significant difference between the two realisations of one phoneme. * = p<0.05, ** = p<0.01, Bold = p<.20

N=51	Intelligibility	Articulation	Nasality	EORTC Speech Scale
Acoustic-Phonetic Analysis				
Vowel space	**.34***	**.29***	**-.24**	**-.19**
/a/1 F1	.14	.06	**-.45****	-.12
/a/1 F2	.13	.08	**-.25**	-.16
/a/1 F2	.13	.08	**-.25**	-.16
/a/2 F1	**.19**	.18	-.11	**-.21**
/a/2 F2	-.15	.10	-.07	.09
/i/1 F1	-.13	**-.23**	**-.33***	.07
/i/1 F2	**.31***	.06	**-.19**	**-.22**
/i/2 F1	.04	-.05	**-.26**	-.03
/i/2 F2	**.20**	**.28***	**-.20**	**-.21**
/u/1 F1	.04	.01	**-.19**	-.15
/u/2 F1	-.05	.02	**-.20**	-.05
/u/ F2 averaged	-.13	-.17	.01	.04
/k/1	**.33***	.18	-.06	**-.24**
/k/2	.00	**-.23**	.14	-.09
/x/1	-.01	**-.19**	.10	.13
/x/2	-.03	.13	**.34***	-.02
/b/ VOT averaged	-.03	**.20**	**-.24**	-.13
/b/1 burst	.11	.13	.02	-.17
/b/2 burst	**-.28***	-.09	-.15	.10
/d/1 VOT	-.01	-.02	.08	.17

/d/2 VOT	.14	**.28**	.10	-.18
/d/ burst averaged	.05	-.04	**.22**	-.07
/p/1 silence	**-.29***	-.04	-.15	**.23**
/p/2 silence	-.07	-.02	-.16	-.17
/p/ burst averaged	.07	-.15	**.25**	**-.21**
/t/ silence averaged	-.03	-.10	.15	-.15
/t/1 burst	.14	.09	**.21**	-.12
/t/2 burst	-.10	.03	.16	.08
Artificial Neural Network Analysis				
ann_nasal	-.18	-.18	-.08	-.06
/a/ ANN averaged	-.04	-.02	.02	-.11
/i/1 ANN	.13	.18	**.38****	-.08
/i/2 ANN	-.17	**-.23**	.02	.04
/u/1 ANN	-.05	-.13	**.21**	.02
/u/2 ANN	-.05	-.18	-.04	-.12
/b/1 VOT ANN	-.01	**-.34***	.03	-.06
/b/1burst ANN	.01	**-.23**	.08	-.11
/b/2 VOT ANN	-.16	**-.31***	.12	-.11
/b/2 burst ANN	-.18	**-.29***	.17	-.13
/d/ VOT ANN averaged	**-.21**	**-.20**	-.02	.08
/d/1 burst ANN	**-.23**	**-.22**	.01	.08
/d/2 burst ANN	**-.22**	-.12	.11	.13
/p/1 silence ANN	.02	-.12	-.12	-.07
/p/1 burst ANN	-.10	-.10	.18	.10
/p/2 silence ANN	-.07	-.13	.14	-.03
/p/2 burst ANN	-.10	**-.20**	**.23**	**-.19**
/t/ silence ANN averaged	-.20	-.17	-.07	.07
/t/1 burst ANN	-.15	**-.23**	.06	.15
/t/2 burst ANN	-.18	-.10	.01	-.04

For cohort 2, both positive and negative Pearson correlations (table 19b) revealed that intelligibility was significantly related to AP analyses of /a, i/ and /x/ and to ANN-voicing of /a/ and /d/. Subjective evaluation of articulation was significantly related to AP analyses of /x/ and to ANN-voicing of /p/. Subjective evaluation of nasal resonance was significantly related to AP analyses of the second formant of /p, t/ and to ANN-nasal on the entire stretch of speech. Subjective evaluation of patient-reported outcome was significantly related to AP analyses of the second formant of /u/ and /t/ and not to ANN-analyses.

Table 19 b. Cohort 2: Pearson correlation coefficient r between objective speech parameters and subjective parameters: intelligibility, articulation, nasality and patient-reported speech outcome. "/a/1" means the first realisation of the two /a/'s that are included in the study, "F1" is the first formant, "F2" is the second formant, "VOT" is the voice-onset-time of stop consonants, "ANN" is analysis by an artificial neural network. Phonemes are "averaged" when there is no significant difference between the two realisations of one phoneme. * = p<0.05, ** = p<0.01, Bold = p<.20

N=64	Intelligibility	Articulation	Nasality	EORTC Speech Scale
Acoustic-Phonetic Analysis				
vowel space	**.22**	**.23**	.11	-.02
/a/1 F1	.03	-.04	.16	-.18
/a/1 F2	**-.44****	-.14	-.15	.09
/a/2 F1	-.07	-.07	-.10	.02
/a/2 F2	-.03	.13	-.16	.14
/i/ F1 averaged	-.08	.13	-.05	.02
/i/1 F2	**.27***	**.23**	.08	-.13
/i/2 F2	**.30***	**.20**	**.23**	**-.22**
/u/1 F1	-.07	-.08	-.05	**-.30***
/u/1 F2	-.01	-.13	.08	-.03
/u/2 F1	.10	-.12	-.01	-.09
/u/2 F2	**-.22**	.01	.01	**-.24**
/x/1	**.27***	**.26***	**.21**	-.03
/x/2	**.32***	**.27***	.10	**-.22**

/k/ averaged	**.25**	.04	.01	−.10
/b/1 VOT	−.07	−.10	.02	.05
/b/2 VOT	.00	−.03	.00	−.11
/b/ burst averaged	−.03	.03	.04	.10
/d/ VOT averaged	.05	−.03	−.08	−.14
/d/ burst averaged	.01	**−.19**	.01	.04
/p/ silence averaged	−.06	.05	−.01	.07
/p/ burst averaged	−.12	−.13	**−.28***	**.19**
/t/1 silence	−.03	−.01	.09	**.32***
/t/2 silence	.07	.14	.13	**−.26**
/t/ burst averaged	.11	−.05	**−.27***	.01
Artificial Neural Network Analysis				
ann_nasal	.15	−.03	**−.26***	.06
/a/1 ANN	**.21**	−.14	−.09	−.04
/a/2 ANN	.02	.01	.02	.05
/i/ ANN averaged	**−.21**	−.15	−.08	.17
/u/ ANN averaged	.05	.09	.15	−.12
/b/1 VOT ANN	**.18**	.01	.02	−.07
/b/2 VOT ANN	.07	−.09	−.10	.15
/b/1 burst ANN	**.17**	−.02	.04	−.06
/b/2 burst ANN	.04	−.14	.00	.17
/d/1 VOT ANN	.14	.12	−.06	−.12
/d/1 burst ANN	.15	.02	−.02	.02
/d/2 VOT ANN	**.37****	.15	−.04	**.22**
/d/2burst ANN	**.35****	.08	−.07	**.24**
/p/1 silence ANN	−.05	**−.27***	−.15	−.05
/p/2 silence ANN	**.23**	−.02	−.13	−.13
/t/1burst ANN	**.22**	−.05	.02	−.16
/t/2 burst ANN	.08	**−.23**	−.13	.09
/t/ ANN silence averaged	.07	−.13	**−.18**	.12

Multivariate regression analyses

To obtain insight into which objective parameters predict significantly subjective speech evaluations, multivariate regression analyses were performed in cohort 1 (table 20a). For the first cohort, the results revealed that intelligibility was predicted by AP analysis of /p/ and vowel space, and by ANN-voicing of /d/. Articulation was predicted by AP analysis of vowel space and by ANN-voicing of /b/. Nasal resonance was best predicted by AP analysis of /a/, /b/ and /x/. Patient reported speech outcome was predicted by AP analysis of /i/ and /k/ and by ANN-voicing of /p/. The amount of variance explained varied from moderate (52.0% for Nasal resonance, 37.7% for Intelligibility and 36.2% for Articulation) to poor (21.1% for patient-reported speech outcome).

Table 20 a. Cohort 1: Prediction of intelligibility, articulation, nasal resonance and patient-reported speech outcome by acoustic-phonetic and Artificial Neural Network analyses. * $p < 0.05$, ** $p < 0.01$.

		b	statistic	p	R^2
Intelligibility	/p/1 silence	-32.21	-3.81	.000**	37.7%
	/d/1 burst ANN	-2.61	-3.05	.004**	
	/d/ VOT ANN averaged	2.32	2.20	.033*	
	Vowel space	3.30	2.10	.045*	
Articulation	vowel space	9.59	5.15	.023*	36.2%
	/b/2 VOT ANN	-4.72	5.27	.022*	
Nasal resonance	/a/1 F1	-0.01	4.81	.028*	52.0%
	/x/2	0.24	7.43	.006**	
	/b/ VOT averaged	-78.38	4.26	.039*	
EORTC Speech Scale	/i/2 F2	-0.01	-2.22	.032*	21.1%
	/k/1	-0.39	-1.71	.095	

Multivariate regression analyses were also performed for cohort 2 (table 20 b). The results revealed that intelligibility was predicted by AP analysis of /a/, /i/ and /x/ and not by ANN analysis. Articulation was predicted by AP analysis of vowel space and by ANN-voicing of /p/. Nasal resonance was best predicted by AP analysis of /t/ and /p/ and not by ANN analysis. Patient reported speech outcome was predicted by AP analysis of /t/ and /u/ and by ANN-voicing of /d/. The amount of variance explained differed from moderate (51.9% for patient-reported outcome and 41.3% for Intelligibility) to poor (21.8% for Nasal resonance and 20.9% for Articulation).

Table 20 b. Cohort 2: Prediction of intelligibility, articulation, nasal resonance and EORTC selfevaluations by acoustic-phonetic and Artificial Neural Network analyses.

* $p < 0.05$, ** $p < 0.01$.		B	statistic	p	R^2
Intelligibility	/a/1 F2	-0.004	-4.16	.000**	41.3%
	/i/ F2	0.002	2.75	.009*	
	/x/ 1	0.06	2.06	.047*	
Articulation	Vowel space	6.96	2.58	.108	20.9%
	/p/1 silence ANN	-2.88	3.04	.082	
Nasal resonance	/p/ burst average	-40.07	3.01	.000**	21.8%
	/t/ burst average	-44.87	3.10	.000**	
EORTC Speech Scale	/t/2 silence	-452.62	-3.66	.001**	51.9%
	/t/1 silence	329.00	3.10	.004**	
	/u/1 F1	-0.10	-2.61	.013*	
	/d/2 burst ANN	20.97	2.25	.031*	

Discussion

This paper presents an inventory of speech quality in a well-defined head and neck cancer patient group six months to nine years after reconstructive surgery and postoperative radiotherapy for advanced oral or oropharyngeal cancer. The aim of the present study was to combine the results from previously performed studies on the development of objective speech analyses[5, 12, 13] to judge the validity of objective speech sounds put together into one model. Firstly, this model was tested in the same cohort of 51 patients as we used in our previous studies.[5, 12, 13] Then a second cohort of 64 patients treated for oral or oropharyngeal cancer was used to externally validate the findings.

Results in the present study confirm our earlier pilot studies.[5, 12, 13] For cohorts 1 and 2 Intelligibility was related to AP analyses of vowel space and /a, i, k, b, p/ and to ANN analyses of /a, d/. Articulation was related to AP analyses of vowel space and /i, x/ and to ANN analyses of /p, b/. Nasal resonance was related to AP analyses of /a, i, x, p, t/ and to ANN analyses of /i/ and ANN-nasal on the entire stretch of speech. Patient-reported outcome was related to AP analyses of vowel space and /u, t/. In both cohorts predictive values remained moderate to poor. Several speech sounds with a p-value <.20 were added to this selection and used in the multivariate model.

In the present study, the multivariate model used in the second cohort of 64 patients required another selection of objective speech analyses than the first cohort of 51 patients, indicating that either both cohorts were not comparable or that the multivariate models are not yet very stable (i.e. dependent on the study cohort). These differences originate from the first methodological step in which it was revealed which two realisations of one speech sound were not normally distributed. If there were significant differences, the two realisations were used separately. The differences in the models show that there is variability in the production of speech sounds and that the models are cohort-dependent. It is possible that a too large amount of variables was used in combination with a relatively small population with large differences in patient characteristics. The model then describes the 'noise' of these parameters and not the underlying patterns, in which case over fitting occurs and predictive value remains low.

The differences between the multivariate models of the two study cohorts may also be explained by demographic and clinical characteristics. Patients in the first cohort were younger (mean 54 vs 60 years) and age has been shown to be a significant factor in voice and speech analyses.[35, 36]
The first cohort involved less patients with smaller T1-T2 tumours (51%) than the second cohort (58%). In the first cohort almost all patients (92%) received radiotherapy, versus 78% in the second cohort. Both tumour stage and treatment modality may have been of influence on speech quality due to the amount of surgically removed tissue and/or stiffening of tissue involved in speech production. It is very well possible that the technique of radiation has been improved over the years between inclusion of patients of both cohorts which causes a difference in clinical characteristics. Also the difference in time between treatment and speech recording (follow-up) was different which may have had influence on the development of fibrosis in the cohort with a longer time until speech recording. In cohort 1, speech recordings were made 6 months after treatment in the first cohort, versus 27 months in the second cohort. Although not investigated, inspection of the recorded speech samples during segmentation gave the impression that speech quality of patients in the second cohort was better compared to speech quality in the first cohort. Also, the scores on the four subjective scales differed significantly and were better for cohort 2 compared to cohort 1. It may be that patients who survived for multiple years (cohort 2) have had more time to adjust to an altered vocal tract and develop strategies to speak more clearly.

Another striking finding in the present study was that, although a higher predictive value was expected now that all speech data as assessed by the two objective methods (AP analyses and ANN analyses) were combined into one model, in both study cohorts the value of objective speech analyses to predict subjective speech evaluation by raters or patients themselves remained moderate at best. In the earlier pilot studies[5, 12, 13] speech sounds were investigated based on classes of speech sounds: phonemes belonging to the class of velar speech sounds, vowels and stop consonants. In the present study these separate phonemes were used concurrently in one model. There may be an interaction between the studied phonemes and other -yet unknown- factors influencing subjective apprehension of speech that could clarify the lower than expected predictive value. Factors that were not yet controlled for but could be taken into account in further research

include a variety of mainly phonetic items. The interaction between phonemes -known as coarticulation and assimilation- is of importance since phonemes are pronounced slightly different due to neighbouring speech sounds. Since phonemes are not pronounced in identical ways it is necessary to use and investigate many different phonological contexts in further research.[37] Second, inspection of word- and sentence composition structure patterns (prosodic categories) is needed because stressed syllables "differ from those that are unstressed along at least four parameters: duration, fundamental frequency, overall intensity, and spectral composition"[38] and thus may influence the dataset. In further research a larger variety of phonemes is therefore needed. Speech sounds belonging to the classes of velar speech sounds, vowels and stop consonants were already investigated in the previous and present studies but could be completed by other phonemes belonging to these classes. Following this expansion other classes of speech sounds could be investigated, such fricatives (/s,z,v,f/), nasals (/m,n,ng/) and trills (/r/). Finally, speaker dependent characteristics should be carefully controlled for, such as influence by personal characteristics (gender, age, health) as well as cultural aspects (sociolect, geographical background, language, accent and dialect of the speaker).[39]

Observing the results and interpretation hereof, it is argued that a speech quality assessment protocol to be used for clinical and research purposes in patients treated for HNC ideally should be multidimensional including subjective, objective and patient-reported speech assessment methods. These seem to be complementary to each other and provide different information. How these sources of information are related to each other is not yet fully understood. Future prospective studies are needed including larger samples of HNC patients.

Conclusion

Combined objective analyses of speech quality in HNC patients by acoustic-phonetic analyses and by Artificial Neural Network analyses are valid and contribute moderately to a multidimensional speech evaluation protocol. More prospective research using more phonemes and larger study samples is needed to improve performance of objective speech quality analysis.

Reference List

1. van der Molen L, van Rossum MA, Burkhead LM, et al. Functional outcomes and rehabilitation strategies in patients treated with chemoradiotherapy for advanced head and neck cancer: a systematic review. Eur Arch Otorhinolaryngol 2009 Jun;266(6):901-2.
2. Borggreven PA, Aaronson NK, Verdonck-de Leeuw IM, et al. Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral Oncol 2007 Nov;43(10):1034-42.
3. Verdonck-de Leeuw I, Borggreven PA, Eerenstein S, et al. Psychosocial and functional consequences of head and neck cancer. Ned Tijdschr Oncol 2006;3(5):185-91.
4. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
5. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop 2009;61(3):180-7.
6. LaBlance GR, Kraus K, Steckol KF. Rehabilitation of swallowing and communication following glossectomy. Rehabil Nurs 1991 Sep;16(5):266-70.
7. Bressmann T, Sader R, Whitehill TL. Consonant intelligibility and tongue motility in patients with partial glossectomy. Journal of Oral and Maxillofacial Surgery 2004;(62):298-303.
8. Sumita YI, Ozawa S, Mukohyama H, et al. Digital acoustic analysis of five vowels in maxillectomy patients. J Oral Rehabil 2002 Jul;29(7):649-56.
9. Michi KI, Imai S, Yamashita Y. Improvement of speech intelligibility by a secondary operation to mobilize the tongue after glossectomy. J Craniofac Surg 1989;17:162-6.
10. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. Folia Phoniatr Logop 2003 May;55(3):147-57.
11. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. J Acoust Soc Am 2007 Oct;122(4):2354-64.
12. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer. Logopedics Phoniatrics Vocology 2011;36(4):168-74.
13. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Acoustic-phonetic and artificial neural network feature analysis to assess speech quality of stop consonants produced by patients treated for oral or oropharyngeal cancer. Speech Communication 2012;54(5):632-40.

14. Bjordal K, Hammerlid E, Ahlner-Elmqvist M, et al. Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. J Clin Oncol 1999 Mar;17(3):1008-19.
15. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365-76.
16. Lee AS, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. Clin Linguist Phon 2003 Jun;17(4-5):259-64.
17. Kent RD. Intelligibility in speech disorders: theory, measurement, and management. Amsterdam, Philadelphia: John Benjamins Publishing; 1992.
18. Bergem Dv. On the perception of acoustic and lexical vowel reduction. Berlin 1993 p. 677-80.
19. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Speech aerodynamics and nasalance in oral cancer patients treated with microvascular transfers. J Craniofac Surg 2005 Nov;16(6):990-5.
20. Tsang CD, Trainor LJ. Spectralslope discrimination in infancy: sensitivity to socially important timbres. Infant behaviour and development 2002;25(2):183-94.
21. Klatt DH. Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. Journal of Speech and Hearing Research 1975;18:686-706.
22. Ladefoged P, I.Maddieson. The sounds of the world's languages. Blackwell Publishing; 1996.
23. Houde J, Jordan M. Sensomotor adaptation in speech production. Science 1998;(279):1213-6.
24. Praat: doing phonetics by computer. [computer program]. Version 5.2.35. University of Amsterdam: 2007.
25. Nabil N, Espy-Wilson CY. A signal representation of speech based on phonetic features. 1995 May 22; Inst. of Tech., Utica/Rome 1995 p. 310-5.
26. Nabil N, Espy-Wilson CY. A knowledge-based signal representation for speech recognition. Atlanta, Georgia 1996 p. 29-32.
27. Deng L, Sun DX. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J Acoust Soc Am 1994;(95):2702-19.
28. Erler K, Freeman GH. An HMM-based speech recognizer using overlapping articulatory features. J Acoust Soc Am 1996;100(4):2500-13.
29. King S, Taylor P. Detection of phonological features in continuous speech using neural networks. Comp Speech Lang 2000;14(4):333-53.
30. Robinson T, Hochberg M, Renals S. The use of recurrent neural networks in contnuous speech recognition. In: Lee C-H, Soong F, ., editors. Automatic

Speech and Speaker Recognition- Advanced Topics. Kluwer Academic Publishers; 1996. p. 233-58.

31. Bridle J, Deng L, Picone J, et al. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. John Hopkins University 1998 p. 1-61.
32. The IFA Spoken Language Corpus. de Nederlandse Taalunie 2001;v.1.
33. Graupe D. Principles of Artificial Neural networks. Advanced series on circuits and systems - volume 6 ed. Singapore: World Scientific Publishing Company Co. Pte. Ltd.; 2007.
34. http://nico.nikkostrom.com/ [computer program]. KTH, Stockholm: 1997.
35. Verdonck-de Leeuw I, Mahieu HF. Vocal aging and the impact on social life: a longitudinal study. Journal of Voice 2004;18(2):193-202.
36. Gorham-Rowan MM, Laures-Gore J. Acoustic-perceptual correlates of voice quality in elderly men and women. J Commun Disord 2006;39(3):171-84.
37. W.J.Hardcastle, N.Hewlett. Coarticulation: Theory, Data And Techniques. New York: Cambridge University Press; 1999.
38. Gay T. Physiological and Acoustic Correlates of Perceived Stress. Language and Speech 1978;21(4):347-53.
39. Schultz T. Speaker Characteristics. Lecture Notes in Computer Science. 4343 ed. 2007. p. 47-74.